Метод порождающих и запрещающих множеств для рандомизированного метода минимизации эмпирического риска*

Фрей А. И.
frey@forecsys.ru
Москва, Московский Физико-Технический Институт (Государственный Университет)

В комбинаторном подходе к проблеме обобщающей способности развиваются методы, позволяющие упрощать вывод точных оценок вероятности переобучения. Для детерминированного метода минимизации эмпирического риска основным инструментом вывода подобных оценок является метод порождающих и запрещающих объектов. В данной работе указанный метод обобщается на случай рандомизированных методов обучения.

При решении задач машинного обучения требуется из заданного множества алгоритмов выбрать алгоритм, который ошибался бы как можно реже не только на объектах наблюдаемой обучающей выборки, но и на объектах скрытой контрольной выборки, которая в момент выбора алгоритма ещё неизвестна. Если частота ошибок на контрольной выборке оказывается значительно выше, чем на обучающей, то говорят, что произошло переобучения алгоритма — он слишком хорошо описывает конкретные данные, но не обладает способностью к обобщению этих данных, не восстанавливает порождающую их зависимость и не пригоден для построения прогнозов.

На практике склонность метода обучения к переобучению оценивается с помощью процедуры скользящего контроля (кросс-валидации). Фиксируется некоторое множество разбиений исходной выборки на две подвыборки — обучающую и контрольную. Для каждого разбиения выполняется настройка алгоритма по обучающей подвыборке, затем оценивается его средняя ошибка на объектах контрольной подвыборки. Оценкой скользящего контроля называется средняя по всем разбиениям величина ошибки на контрольных подвыборках.

Нашей целью будет получение оценок функционала скользящего контроля без применения процедуры кросс-валидации.

Основные обозначения

Пусть задана генеральная выборка $\mathbb{X} = (x_1, \ldots, x_L)$, состоящая из L объектов. Произвольный алгоритм классификации, примененный к данной выборке, порождает бинарный вектор ошибок $a \equiv \big(I(a, x_i)\big)_{i=1}^{L}$, где $I(a, x_i) \in \{0, 1\}$ — индикатор ошибки алгоритма a на объекте x_i. В дальнейшем алгоритмы будут отождествляться с векторами их ошибок на выборке $\mathbb{X}$.

Обозначим через $\mathbb{A} = \{0, 1\}^L$ множество всех возможных векторов ошибок длины L. Через $[\mathbb{X}]^\ell$ обозначим множество всех разбиений генеральной выборки $\mathbb{X}$ на обучающую выборку X длины ℓ и контрольную выборку $\bar{X}$ длины $k = L - \ell$. *Число ошибок* алгоритма a на выборке $U \subseteq \mathbb{X}$ обозначим через $n(a, U) = \sum_{x \in U} I(a, x)$. Величину $\nu(a, U) = n(a, U)/|U|$ будем называть *частотой ошибок* алгоритма a на выборке U. *Уклонение частот* на разбиении $\mathbb{X} = X \sqcup \bar{X}$ определим как разность частот ошибок на контроле и на обучении: $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$.

Пусть $A \subset \mathbb{A}$ — множество алгоритмов с попарно различными векторами ошибок. Обозначим через $A(X)$ множество алгоритмов с минимальным числом ошибок на обучающей выборке X:

$$A(X) = \operatorname*{Argmin}_{a \in A} n(a, X). \qquad (1)$$

Частоту ошибок на обучающей выборке называют *эмпирическим риском*. *Минимизация эмпирического риска* μ — это метод обучения, который из заданного множества $A \subset \mathbb{A}$ выбирает алгоритм $a \in A$, допускающий наименьшее число ошибок на обучающей выборке X. Таким образом, для всех $X \in [\mathbb{X}]^\ell$ выполнено $\mu X \in A(X)$.

Говорят, что метод μ переобучен на разбиении $X \sqcup \bar{X}$, если уклонение частот $\delta(a, X)$ превышает фиксированный порог ε. Переобучение может быть следствием «неудачного» разбиения генеральной выборки на обучение и контроль. Поэтому вводится функционал *вероятности переобучения*, равный доле разбиений выборки, при которых возникает переобучение [1, 2]:

$$Q_\varepsilon(A) = \mathsf{E}[\delta(\mu X, X) \geqslant \varepsilon], \text{ где } \mathsf{E} = \frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell}.$$

Тут и далее квадратные скобки — нотация Айверсона, переводящая логическое выражение в число 0 или 1 по правилам [истина] = 1, [ложь] = 1.

Функционал $Q_\varepsilon(A)$ уже не зависит от выбора разбиения и характеризует качество данного метода обучения на данной генеральной выборке.

Работа поддержана РФФИ (проект № 08-07-00422) и программой ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Теорема о порождающих и запрещающих объектах

Первый подход, позволивший получать точные оценки вероятности переобучения в рамках слабой вероятностной аксиоматики, основан на выделении порождающих и запрещающих объектов [3].

Гипотеза 1. Пусть множество A, выборка $\mathbb{X}$ и детерминированный метод обучения μ таковы, что для каждого алгоритма $a \in A$ можно указать пару непересекающихся подмножеств $X_a \subset \mathbb{X}$ и $X'_a \subset \mathbb{X}$, удовлетворяющую условию

$$\big[\mu X{=}a\big] = \big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big], \quad \forall X \in [\mathbb{X}]^\ell. \quad (2)$$

Множество X_a будем называть *порождающим*, множество X'_a — *запрещающим* для алгоритма a. Гипотеза 1 означает, что метод μ выбирает алгоритм a тогда и только тогда, когда в обучающей выборке X находятся все порождающие объекты и ни одного запрещающего. Все остальные объекты $\mathbb{X}\backslash X_a\backslash X'_a$ будем называть *нейтральными* для алгоритма a.

Для произвольного $a \in A$ обозначим через L_a число нейтральных объектов, через ℓ_a — число нейтральных объектов, попадающих в обучающую выборку:

$$L_a = L - |X_a| - |X'_a|;$$
$$\ell_a = \ell - |X_a|.$$

Теорема 1. *Если гипотеза 1 справедлива, то вероятность получить в результате обучения алгоритм a равна*

$$P_a(A) = \mathsf{P}[\mu X{=}a] = \frac{C^{\ell_a}_{L_a}}{C^\ell_L},$$

а вероятность переобучения $Q_\varepsilon(A)$ выражается по формуле полной вероятности:

$$Q_\varepsilon(A) = \sum_{a\in A} P_a\,\mathsf{P}\big(\delta(a,X) \geqslant \varepsilon \mid a\big) = \sum_{a\in A} P_a H^{\ell_a,m_a}_{L_a}\left(s_a(\varepsilon)\right).$$

Данный результат позволил получить формулы вероятности переобучения для широкого класса модельных семейств алгоритмов, в частности для монотонных и унимодальных сеток. В следующем параграфе аналогичная теорема будет получена для рандомизированных методов обучения.

Рандомизированный метод минимизации эмпирического риска

Рандомизированный метод минимизации эмпирического риска выбирает произвольный алгоритм из множества $A(X)$ случайно и равновероятно [4, 5].

Поскольку в задаче статистического обучения появляется второй независимый источник случайности, определение вероятности переобучения $Q_\varepsilon(A)$ приходится модифицировать. Наиболее естественный вариант модификации — усреднение по множеству $A(X)$:

$$Q_\varepsilon(A) = \mathsf{E}\frac{1}{|A(X)|}\sum_{a\in A(X)}\big[\delta(a,X) \geqslant \varepsilon\big]. \quad (3)$$

Для детерминированного метода обучения результатом обучения являлся алгоритм $a \in A$. В случае рандомизированного метода обучения результатом обучения является подмножество $A(X) \subset A$. Данное обстоятельство позволяет сформулировать следующую гипотезу.

Гипотеза 2. Пусть $\aleph = \{A(X) : X \in [\mathbb{X}]^\ell\}$ — множество всех $A(X)$, получающихся в результате обучения. Пусть множество A и выборка $\mathbb{X}$ таковы, что для каждого алгоритма $\alpha \in \aleph$ можно указать пару непересекающихся подмножеств $X_\alpha \subset \mathbb{X}$ и $X'_\alpha \subset \mathbb{X}$, удовлетворяющую условию

$$\big[A(X){=}\alpha\big] = \big[X_\alpha \subseteq X\big]\big[X'_\alpha \subseteq \bar{X}\big], \quad \forall X \in [\mathbb{X}]^\ell. \quad (4)$$

Теорема 2. *Если гипотеза 2 справедлива, то вероятность переобучения $Q_\varepsilon(A)$ выражается по следующей формуле:*

$$Q_\varepsilon(A) = \sum_{a\in A}\sum_{\alpha\in\aleph}\frac{[a\in\alpha]}{|\alpha|}\frac{C^{\ell_\alpha}_{L_\alpha}}{C^\ell_L}H^{\ell_\alpha,m^a_\alpha}_{L_\alpha}(s^a_\alpha(\varepsilon)),$$

где $L_\alpha = L - |X_\alpha| - |\bar{X}_\alpha|$, $\ell_\alpha = \ell - |X_\alpha|$, $m^a_\alpha = n(a,\mathbb{X}\backslash X_\alpha\backslash X'_\alpha)$, $s^a_\alpha(\varepsilon) = \frac{\ell}{L}(n(a,\mathbb{X}) - \varepsilon k) - n(a,X_\alpha)$, $H^{\ell,m}_L(z) = \sum_{s=0}^{\lfloor z\rfloor}\frac{C^s_m C^{\ell-s}_{L-m}}{C^\ell_L}$ — левый хвост гипергеометрического распределения.

Лемма 3. *Пусть в множестве A есть алгоритм a_0, такой что для любого $a \in A$ вектор ошибок алгоритма a_0 содержится в векторе ошибок алгоритма a. Обозначим множество объектов на которых ошибается a_0 через X_0. Пусть система порождающих и запрещающих множеств такова, что для всех α выполнено $X_0 \cap X_\alpha = \varnothing$ и $X_0 \cap X'_\alpha = \varnothing$. Тогда в формуле порождающих и запрещающих объектов можно следующим образом упростить обозначения: $m^a_\alpha = n(a_0,\mathbb{X})$, $s^a_\alpha(\varepsilon) = \frac{\ell}{L}(n(a,\mathbb{X}) - \varepsilon k)$.*

Теорема о порождающих и запрещающих объектах легко объединяется с теоремой о разбиении множества алгоритмов на орбиты [4].

Лемма 4. *Пусть $G \subset \operatorname{Sym} A$ — подгруппа группы симметрии множества алгоритмов A, $\Omega(A)$ — множество всех орбит действия G на A, a_ω — произвольный представитель орбиты $\omega \in \Omega$, а остальные обозначения — как в теореме 2. Тогда вероятность*

переобучения $Q_\varepsilon(A)$ можно записать в виде:

$$Q_\varepsilon(A) = \sum_{\omega\in\Omega(A)} \sum_{\alpha\in\aleph} [a_\omega \in \alpha] \frac{|\omega|}{|\alpha|} \frac{C^{\ell_\alpha}_{L_\alpha}}{C^{\ell}_{L}} H^{\ell_\alpha, m^{a_\omega}_\alpha}_{L_\alpha}(s^{a_\omega}_\alpha(\varepsilon)).$$

Гипотеза 2 накладывает слишком сильные ограничения на выборку $\mathbb{X}$ и семейство A. Рассмотрим естественное обобщение: предположим, что для каждого алгоритма a существуют различные варианты выделения порождающих и запрещающих множеств.

Гипотеза 3. Пусть множество A и выборка $\mathbb{X}$ таковы, что для каждого алгоритма $\alpha \in \aleph$ можно указать конечное множество индексов V_α, и для каждого индекса $v \in V_\alpha$ можно указать *порождающее* множество $X_{\alpha v} \subset \mathbb{X}$, *запрещающее* множество $X'_{\alpha v} \subset \mathbb{X}$ и коэффициент $c_{\alpha v} \in \mathbb{R}$, для всех $X \in [\mathbb{X}]^\ell$ удовлетворяющие условиям

$$[A(X){=}\alpha] = \sum_{v\in V_a} c_{\alpha v} [X_{\alpha v} \subseteq X] [X'_{\alpha v} \subseteq \bar{X}]. \quad (5)$$

При условиях данной гипотезы теорема 2 о порождающих и запрещающих множествах обобщается полностью аналогично случаю детерминированного метода обучения.

Важно отметить, что гипотеза 3 выполнена для произвольной выборки $\mathbb{X}$ и множества алгоритмов A.

Теорема 5. *Для любых $\mathbb{X}$ и A существуют множества V_α, $X_{\alpha v}$, $X'_{\alpha v}$, при которых справедливо представление* (5)*, причём $c_{\alpha v} = 1$ для всех $\alpha \in A$, $v \in V_\alpha$.*

Монотонные и унимодальные сети алгоритмов

Монотонная сетка алгоритмов [6] — это модель параметрического *связного семейства алгоритмов*, предполагающая, что при непрерывном удалении каждой компоненты вектора параметров от оптимального значения число ошибок на полной выборке только увеличивается.

Пример 1. Монотонная двумерная сетка при $m = 0$ и $L = 4$:

$$\begin{array}{c} x_1 \\ x_2 \\ \hline x_3 \\ x_4 \end{array} \left(\begin{array}{ccccccccc} 0 & \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & \mathbf{1} \\ 0 & 0 & \mathbf{1} & 0 & 0 & \mathbf{1} & 0 & 0 & \mathbf{1} \\ \hline 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} \end{array}\right)$$

Унимодальная сетка алгоритмов [6] является более реалистичной моделью связного параметрического семейства, по сравнению с монотонной сеткой. Если мы имеем лучший алгоритм a_0 с оптимальным значением вектора вещественных параметров, то отклонение значений компонент этого вектора как в бóльшую, так и в меньшую, сторону приводит к увеличению числа ошибок.

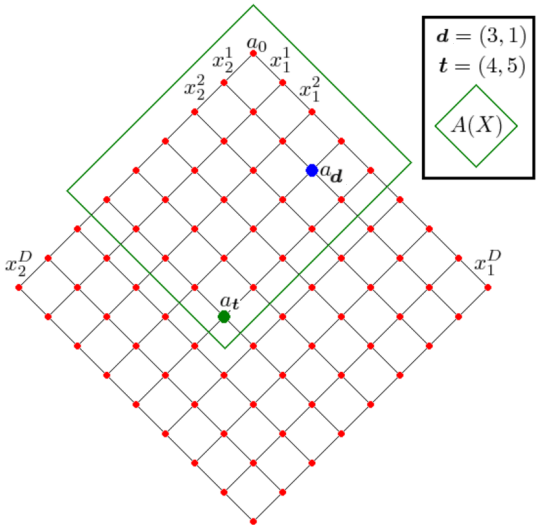

Рис. 1. Строение множества $A(X)$ для двумерной монотонной сетки; $h = 2$, $D = 8$.

Формулы вероятности переобучения рандомизированного метода минимизации эмпирического риска для многомерной монотонной и унимодальной сеток уже были получены ранее [7]. Однако ранее не отмечалось, что данные множества удовлетворяют условиям гипотезы 2 о порождающих и запрещающих объектах для рандомизированного метода минимизации эмпирического риска. Данное обстоятельство позволяет упростить вывод точных формул вероятности переобучения.

Введём целочисленный вектор индексов $\boldsymbol{d} = (d_1, \ldots, d_h) \in \mathbb{Z}^h$. Обозначим $\|\boldsymbol{d}\| = \max\limits_{j=1,\ldots,h} |d_j|$, $|\boldsymbol{d}| = |d_1| + \ldots + |d_h|$. На множестве векторов индексов введём покомпонентное отношение сравнения: $\boldsymbol{d} < \boldsymbol{d}'$, если $d_j \leqslant d'_j$, $j = 1, \ldots, h$, и хотя бы одно из неравенств строгое.

Определение 1. *Множество алгоритмов $A = \{a_{\boldsymbol{d}}\}$, где $\boldsymbol{d} \geqslant 0$ и $\|\boldsymbol{d}\| \leqslant D$ называется монотонной h-мерной сеткой алгоритмов длины D, если существует $h \in \mathbb{N}$ и упорядоченные наборы объектов $X_j = \{x_j^1, \ldots, x_j^D\} \subset \mathbb{X}$, для всех $j = 1, \ldots, h$, а так же множества $U_1 \subset \mathbb{X}$ и $U_0 \subset \mathbb{X}$, такие что:*

1. *набор $\left\{U_0, U_1, \{X_j\}_{j=1}^h\right\}$ является разбиением множества $\mathbb{X}$ на непересекающиеся подмножества;*
2. *$a_d(x_j^i) = [i \leqslant d_j]$, где $x_j^i \in X_j$;*
3. *$a_d(x_0) = 0$ при всех $x_0 \in U_0$;*
4. *$a_d(x_1) = 1$ при всех $x_1 \in U_1$.*

Следующая лемма утверждает, что множество алгоритмов A удовлетворяет условиям гипотезы 2.

Лемма 6. *Рассмотрим произвольное $\alpha \in \aleph$. Согласно определению $\aleph$ это значит, что найдется такое разбиение $X \in \mathbb{X}$, для которого $\alpha = A(X)$.*

Обозначим $i'(j)$ наименьший номер $i \in 1,\ldots,D$, такой что объект x_j^i попадает в обучающую выборку X. Возможно, что для некоторых $j = 1,\ldots,h$ выражение $i'(j)$ не определено, поскольку все объекты x_j^i оказались в контроле $\bar{X}$. Пусть $J \subset 1,\ldots,h$ — множество индексов j, для которых $i'(j)$ определено.

Тогда $X_\alpha = \bigcup\limits_{j\in J} x_j^{i'(j)}$, $X'_\alpha = \bigcup\limits_{j=1}^{h} \bigcup\limits_{i=1}^{i'_j-1} x_j^i$, причем построенные таким образом множества X_α и X'_α зависят только от исходного множества α, но не от выбора представителя $X \in \mathbb{X}$, такого что $\alpha = A(X)$.

Данная лемма в сочетании с теоремой 2 позволяет выписать формулу вероятности переобучения рандомизированного метода обучения для многомерной монотонной сетки алгоритмов.

Теорема 7. *Вероятность переобучения рандомизированного метода минимизации эмпирического риска, примененного к монотонной сетке $A = \{a_{\boldsymbol{d}}\}$ размерности h, $\|\boldsymbol{d}\| \leqslant D$, дается выражением:*

$$Q_\varepsilon(A) = \sum_{\substack{\boldsymbol{d}\geqslant\boldsymbol{0},\\ \|\boldsymbol{d}\|\leqslant\boldsymbol{D}}} \sum_{\substack{\boldsymbol{t}\geqslant\boldsymbol{0},\\ \|\boldsymbol{t}\|\leqslant\boldsymbol{D}}} \frac{[\boldsymbol{t}\geqslant\boldsymbol{d}]}{V(\boldsymbol{t})} \frac{C_{L'}^{\ell'}}{C_L^\ell} H_{L'}^{\ell',m}(s(\varepsilon)),$$

$V(\boldsymbol{t}) = \prod_j (t_j+1)$, $\ell' = \ell - \sum_{j=1}^h [t_j \neq D]$, $k' = k - |\boldsymbol{t}|$, $L' = \ell' + k'$, $s(\varepsilon) = \frac{\ell}{L}[m + |\boldsymbol{d}| - \varepsilon k]$.

Выводы

В данной работе удалось обобщить метод порождающих и запрещающих объектов на случай рандомизированных методов обучения. Полученный результат по прежнему позволяет учитывать структуру симметрии множества алгоритмов (разбиение множества алгоритмов на орбиты действия группы симметрии).

Как и для случая детерминированных методов обучения, доказано что для любого множества алгоритмов можно указать систему порождающих и запрещающих множеств. К сожалению, данный результат вновь является типичной теоремой существования: использованный при её доказательстве способ построения множества порождающих и запрещающих множеств требует явного перебора всех разбиений выборки, что приводит к вычислительно неэффективным оценкам вероятности переобучения.

Тем не менее, с помощью предложенного подхода получены эффективные оценки вероятности переобучения для модельных семейств алгоритмов: монотонных и унимодальных многомерных сеток.

Литература

[1] *Воронцов К. В.* Точные оценки вероятности переобучения // Доклады РАН, 2009. — Т. 429, № 1. — С. 15–18.

[2] *Воронцов К. В.* Комбинаторный подход к проблеме переобучения // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009. — С. 18–21.

[3] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, no. 3. — Pp. 269–285.

[4] *Фрей А. И.* Точные оценки вероятности переобучения для симметричных семейств алгоритмов // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009. — С. 66–69.

[5] *Frei A. I.* Accurate estimates of the generalization ability for symmetric set of predictors and randomized learning algorithms // Pattern Recognition and Image Analysis. — 2010. — Vol. 20, no. 3. — Pp. 241–250.

[6] *Ботов П. В.* Точные оценки вероятности переобучения для монотонных и унимодальных семейств алгоритмов // Всеросс. конф. ММРО-14 — М.: МАКС Пресс, 2009. — С. 7–10.

[7] *Фрей А. И.* Вероятность переобучения плотных и разреженных многомерных сеток алгоритмов // Интеллектуализация обработки информации — М.: МАКС Пресс, 2010. — С. ??–??.